PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Electroencephalography as a Clinical Tool for Diagnosing and Monitoring Attention Deficit Hyperactivity Disorder – A Cross Sectional Study
AUTHORS	Helgadóttir, Halla; Gudmundsson, Olafur; Baldursson, Gísli; Magnússon, Páll; Blin, Nicolas; Brynjólfsdóttir, Berglind; Emilsdóttir, Ásdís; Gudmundsdóttir, Gudrún; Kristmundsdóttir, Kristín; Lorange, Málfrídur; Newman, Paula; Jóhannesson, Gísli; Johnsen, Kristinn

VERSION 1 - REVIEW

REVIEWER	Dr Lars Michels Department of Neuroradiology University Hospital of Zurich
REVIEW RETURNED	18-Jul-2014

GENERAL COMMENTS	1) Please provide a table with a summary of the demographical data (e.g. age ranges of all groups, mean and range of DSM-IV scores, medication, comorbidities). Also, I would like to know how many of many of the combined subtype were boys? The exclusion criterion was severe intellectual disability. What does this mean? Did the authors run IQ subtest? Do they have data on this, which could be added to Table 1? 2) Do the authors have a measurement for puberty? In other words, can be believe that all children can be considered as children, or did some of them already turned adolescent? 3) How many of the subjects were medicated? Can the analysis be repeated excluding the medicated children (given the large (236) number of children without treatment -> page 8)? Or can medication (e.g. dose) be treated as a covariate of no interest in the classification analysis, as we know that medication affects the EEG of ADHD subjects (Clarke et al. 2007; Barry et al., 2009)? Actually, the authors have the chance to provide classification results for both the treated and untreated sample. 4) The classification rate is relatively low, given the nice large sample size. How does the classification result change if authors only include the combined subtype (which covers 64% their data) or just the inattentive type (34%)? So far the heterogeneity is considered to be a limitation and as a strength (page 16). I fully agree with the later, if additional classification analysis will be repeated with respect to the

	ADHD subtype. 5) A big plus of this study is that the authors have independent data to validate their classification algorithms. But age and gender could be provided (Table 1) for the data on the 35 ADHD and 36 control recordings (e.g. do we see the gender bias?). More importantly, what kind of subtype was dominant for these 35 ADHD subjects? If they are mainly of the combined subtype, it might be worth (see (6)) to repeat (all) classification analyses just with the combined-type subjects. 6) I found it surprising that 22% of the subjects had two or more comorbid disorders. What type? Which group was dominantly affected, e.g. the combined subtype? Due to the large nr. of subjects, the authors could repeat the classification analysis removing this 22% of data. 7) Introduction: There are a bunch of classification studies using power values apart from Magee et al (2005). In this paper, specificity was only 40% but other papers (e.g. Poil et al., 2014) found much better classification rates (> 70%). Authors could cite other studies, using classification analyses, to balance the introduction. 8) MAJOR: So far “coherence” or “coherence features” (page 16) remain loose definitions. The authors need to review some of the existing (ADHD) coherence studies in more detail (e.g. papers by Dupuy). What kind of coherence measures were used in those studies (references 21 and 22 but only in other studies)? Which (scalp/brain) regions demonstrate abnormal connectivity in ADHD? Critically, what kind of coherence (for which frequency band) was used in the present study (lagged coherence, zero-lag coherence, global synchronization)??? The information provided on page 9 is not sufficient. 9) Although the ADHD rating scale IV is a validated scale, the rationale of using a “score of 1.5 standard deviations above the age-appropriate norm” needs to be justified. 10) Methods: How much clean (after artefact correction) EEG data was finally obtained in the examined groups? This is important to know since the overall recording duration was only three minutes. 11) Major 1: I am missing pure coherence results. Do the groups (especially in the four age sub-groups) differ in coherence in specific frequency-bands? So far, it is unclear whether the classification results really support the findings by Magee et al (2005) and Barry et al. (2002). See also comment (15). 12) Major 2: The authors have data on power values as well (Table 3). What was the reason to not include them into the construction of the classifiers? Would this not lead to better classification results? In general it would be interesting to see if power-related classification results differ to coherence-related classification results. 13) Why the authors did use nine (age-independent) classifiers for the construction (page 10, bottom)?

	14) Can the authors discuss the results of Table 5 (and Table 7) in more detail, as the highest accuracy and AUC values was achieved for age group 4? Could this be the result of a developmental lag, which is present in ADHD population (whereas the non-ADHD subjects already moved to puberty)? 15) Also the results are not discussed in the context of other EEG coherence studies (to ADHD). 16) Finally, a clinician will not necessarily know what to do with these results. First, the classification results are moderate (< 80%). Second, the results are averaged across different sub-types of ADHD, and third, medication was not talking into account so far. As already written, authors can address issues (2) and (3) by more detailed analysis. 17) Although a nice try, authors do not have any support for the assumption that their connectivity results are driven/linked by alterations in the neurotransmitter (e.g. dopaminergic) system. In fact, results could be partially driven medication, which affects neurotransmission, and ultimately coherence. 18) Only because all recordings were made in Iceland, does not allow saying: "genetics…were quite tightly controlled". Please rephrase (e.g. page 16).

REVIEWER	Sandra Loo and Iman M. Rezazadeh University of California, Los Angeles USA
REVIEW RETURNED	29-Jul-2014

GENERAL COMMENTS	I'm not sure about this. Statistical pattern recognition is used to determine the classifier. I am not an expert in the use of this algorithm. The goal of this study is to develop a multivariate EEG biomarker for the diagnosis of ADHD. The authors report 72-79% diagnostic accuracy for ADHD and conclude this may be useful for improving accuracy of initial diagnosis and ongoing treatment monitoring of children with ADHD. Overall, there are several strengths of this manuscript. The paper is clearly written and includes a very large sample. The use of EEG coherence is novel and sophisticated statistical techniques (statistical pattern recognition; SPR) are used in determining the classifier. Finally, an independent sample of EEGs was used to test the classifier. There are several weaknesses, however, particularly with regard to study methodology, data analysis, and scientific significance that diminish the impact of this manuscript. The lack of detail makes it difficult to review the study's methodological rigor and the omission

	of the measures that are retained for the final classifier makes evaluation of the scientific contribution and impact impossible. Specific details that would make the report of this study more complete and comprehensive follow. With regard to study methodology, more detail and better characterization of the ADHD sample is needed. The only exclusion criterion for the ADHD sample was 'moderate to severe intellectual disability', what IQ range does this include? Was the intelligence measured in the control sample and the same exclusion criterion applied? Further, it is noted that patients with co-morbid autism spectrum disorder (ASD) were included and reported to be a frequent co-morbidity. All co-morbidities and their frequencies should be reported in the paper. Because ASD is often so much more severe than ADHD, it is questionable whether subjects with ASD should have been included in the sample. The authors should report whether ASD or any other co-morbidity affected the diagnostic accuracy of the classifier. The control sample seems to have undergone minimal evaluation. The threshold of <1.5 SD below the age norm seems too liberal for a control sample and may have included controls who have sub-threshold ADHD. Authors should report behavior rating scores for control subjects who were misclassified in the analysis. Overall, more demographic (SES, race, gender) and clinical (comorbidities, ADHD severity and symptoms, IQ, behavioral functioning) data should be reported for both samples, ADHD and control, so that generalizability of the data can be assessed. Data processing and analysis: Because there were 295 ADHD patients and 350 recordings, there are some recordings that were made on the same person during on and off medication conditions, as explained in the manuscript. Presumably, these are considered 'independent' even though recordings from the same person were likely included in the training and testing cohorts. Since the testing cohort only includes 35 EEGs, it could very well be that all of the EEGs were from people also used in the training phase. This calls into question just how 'independent' the testing cohort really was. Although there may have been some EEG changes while on and off medications, the EEGs are likely highly correlated and unlikely to be completely independent in the training and testing cohorts. This contradicts the authors statements that the testing of the classifier was done in an independent sample. It is not clear how authors dealt with the variation among the recording sites, even though similar systems and settings were used, before combining the EEG data and training the classifier. More details should be given regarding how the EEG data were recorded and processed. Please provide information on how the data were cleaned of artifacts and noise. What was the minimum number of epochs needed to be included in the analysis? Were 100% of the EEGs obtained on the ADHD sample usable and included in the analysis? How were the power spectrum, relative power and coherence calculated? How was volume conductance

	accounted for when calculating coherence? For the statistical analysis, a more complete description of the rationale of the SPR classification method/classifier should be given; why were other widely used classifiers like Support Vector Machine, ANN, and ANFIS not used? The 'training' set has near 90% of the data, which is more than what is normally suggested for training a classifier. Please give the rationale for this rather than using more standard percentages, e.g., splitting the dataset around 70% 'training set', 20% 'test set' and 10% 'validation set'. What is the main reason/s that accuracy of age-dependent /independent classifiers differs? Is this related to sample size or other factors? What is the mutual information/cross-correlation of the coherence features that are used in the analysis? It is not clear in the paper why subsets of 20-features have been chosen for the evolutionary classification method? Why 20? How was the redundancy among the 20-set of features reduced? What is the final subset of features used in the classifiers? Please list the features. Did the authors test the association between the final features and clinical features of the sample? This is an important step to determine the functional significance of the features used in the classifier. The omission of the results also makes it impossible to compare to previous studies that have examined coherence measures in ADHD as well as previous findings on the neurobiology of ADHD. Diagnostic accuracy of ~80% leaves approximately 20% of subjects incorrectly diagnosed. Authors should outline the appropriate usage of an algorithm with only moderate diagnostic accuracy.

REVIEWER	Victoria Harris King's College London, United Kingdom
REVIEW RETURNED	05-Sep-2014

GENERAL COMMENTS	This is primarily a statistical review. The paper explores an interesting application of classification tools to EEG data in order to identify individuals with ADHD. The authors use ROC analysis, comparing sensitivity, specificity and area under the curve, to assess the performance of age specific and non-age specific classifiers. The research is novel in that although similar methods have been applied successfully to other disorders, this is the first study for which these methods have been applied to a group with ADHD. This suggests that this technique may have application as a diagnostic tool for clinicians. Overall the analysis appears sound but it would be useful to the reader for the authors to supply additional details on the classifiers and how they were constructed. The authors mention that they divide their data in to a training cohort of 315 cases and the same number of controls and a test cohort of 35 cases and 36 controls. They mention that 350 of the case observations were obtained from 295 individuals. Were any of the individuals that were measured more than once included in both the

	test and the training set and if so did you account for this? On page 10, lines 4-11 the authors describe how they selected their classifiers using statistical pattern recognition and an evolutionary algorithm based on the target of the AUC. Statistical pattern recognition refers to a set of statistical learning techniques, in this case for the purpose of classification. It would be helpful to have further detail here of the SPR technique used to allow the reader to better understand how the candidate classifiers have been constructed. An evolutionary algorithm is an optimisation technique that models candidate solutions as individuals in a population, modelling recombination of features within each iteration, and here is used to select a set of twenty features associated with each classifier. It would be beneficial to add a brief description of the method and some additional references. In the discussion the authors mention the possible issue of confounding due to gender. They also mention that a large proportion (66%) had comorbid disorders. The authors also mention that SPR of EEG data has previously been successfully used for data on autism. How do the classifiers for autism found in previous studies compare with those for ADHD for in this study? Similarly has any previous work explored gender differences in EEG readings?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name **Dr Lars Michels**
Institution and Country **Department of Neuroradiology**
University Hospital of Zurich
Please state any competing interests or state 'None declared': **None declared**

1.1) Please provide a table with a summary of the demographical data (e.g. age ranges of all groups, mean and range of DSM-IV scores, medication, comorbidities). Also, I would like to know how many of many of the combined subtype were boys? The exclusion criterion was severe intellectual disability. What does this mean? Did the authors run IQ subtest? Do they have data on this, which could be added to Table 1?

A table of demographical data including age, gender and comorbidities for each subtype of ADHD. To explain the severity of ADHD symptoms, we include the scores of the ADHD rating scale, see table 1b. For medication see table 2a and table 2b.

Table 1. ADHD subtypes: Comorbidity and gender.

		total	male	female	Mean age	1 comorbid	2 comorbid
ADHD group		310	238	72	9,58	185	66
	ADHD IA	102	77	25	9,93	60	23
	ADHD COM	202	155	47	9,45	124	43
	ADHD H/I	6	6	0	8,97	1	0
controls		351	175	176	9,53	0	0

Table 1b. ADHD training cohort. ADHD rating scale results.

		total	Score IN	Score H/I	Score TS	Score SD above mean

ADHD group		274	14,97	18,85	33,81	2,85
	ADHD IA	89	11,31	18,33	29,64	2,52
	ADHD COM	180	16,85	19,25	36,10	3,05
	ADHD H/I	5	14,50	14,25	28,75	1,83
controls		315	3,75	2,31	6,07	-0,29

Score IA : ADHD rating scale, inattentive score
Score H/I : ADHD rating scale, hyperactivity and impulsiveness score
Score TS : ADHD rating scale, total score
ADHD SD above: Standard deviations above average on ADHD rating scale

IQ tests were not conducted for the normal comparison group but moderate and severe intellectual disability was excluded by parent information about mental or developmental problems. Tests of intellectual level were available in the two clinics from which the ADHD participants were recruited, making it possible to exclude children with moderate and severe intellectual disability (IQ<50). IQ scores were not recorded in the dataset of the study.

1.2) Do the authors have a measurement for puberty? In other words, can be believe that all children can be considered as children, or did some of them already turned adolescent?

Response: It is indeed possible that some of the subjects have already turned adolescent, therefore we have corrected parts the manuscript related to our dataset and changed "children" to "children and adolescents", "subjects" or "patients".

1.3) How many of the subjects were medicated? Can the analysis be repeated excluding the medicated children (given the large (236) number of children without treatment -> page 8)? Or can medication (e.g. dose) be treated as a covariate of no interest in the classification analysis, as we know that medication affects the EEG of ADHD subjects (Clarke et al. 2007; Barry et al., 2009)? Actually, the authors have the chance to provide classification results for both the treated and untreated sample.

Response: See general response, we want both groups to be included. No difference in ADHD medicated vs. ADHD not medicated was found using the ADHD vs. control index. Therefore medication does not affect the classifiers. We are very interested in how treatment affects EEG and we intend to research this effect, but it is beyond the scope of this paper. The ROC curve results have been placed in a new table in the paper. The corresponding ROC curve is shown here in figure 1.

Table 2a. ADHD medication: training cohort

		total	2 visits	Medication	Methylphenidate	Atomoxetine	Meth+Atom
ADHD group		274	41	100	76	22	2
	adhd ia	89	18	35	28	7	0
	adhd com	180	23	65	50	15	0
	adhd hyp	5	0	0	0	0	0

Table 2b. ADHD medication: independent cohort

		total	2 visits	Medication	Methylphenidate	Atomoxetine	Meth+Atom
ADHD group		36	0	10	8	2	0
	adhd ia	13	0	2	4	0	0
	adhd com	22	0	8	4	2	0
	adhd hyp	1	0	0	0	0	0

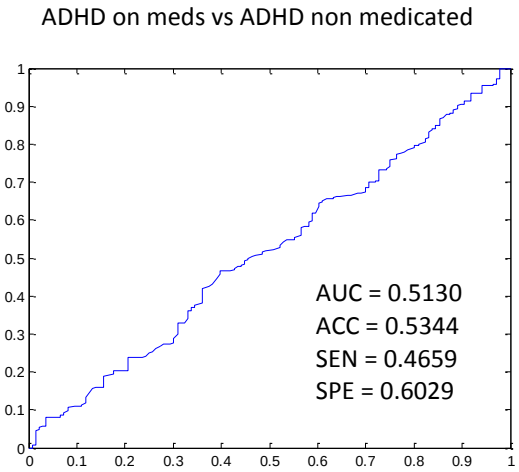


Figure 1. The ROC curve for the separation of ADHD medicated vs. not medicated, based on the ADHD vs. control index. The AUC value is close to 0.5 and thus there is no separation.

1.4) The classification rate is relatively low, given the nice large sample size. How does the classification result change if authors only include the combined subtype (which covers 64% their data) or just the inattentive type (34%)? So far the heterogeneity is considered to be a limitation and as a strength (page 16). I fully agree with the later, if additional classification analysis will be repeated with respect to the ADHD subtype.

Response: See general response, we want both groups to be included. No difference in ADHD COM vs ADHD IA was found using the ADHD vs control index. Therefore ADHD subtype does not affect the classifiers. We intend to investigate the difference between ADHD subtypes using this methodology, but it is beyond the scope of this paper. The ROC curve results have been placed in a new table in the paper. The corresponding ROC curve is shown here in figure 2.

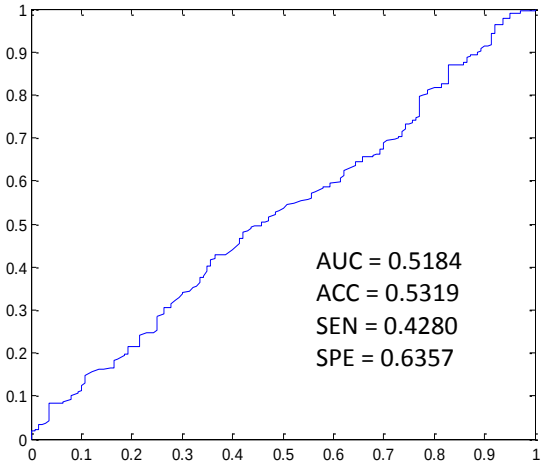


Figure 2. The ROC curve for the separation of ADHD combined type (N=264) vs. ADHD inattentive type (N=140) based on the ADHD vs. control index. The AUC value is close to 0.5 and thus there is no separation.

1.5) A big plus of this study is that the authors have independent data to validate their classification algorithms. But age and gender could be provided (Table 1) for the data on the

35 ADHD and 36 control recordings (e.g. do we see the gender bias?). More importantly, what kind of subtype was dominant for these 35 ADHD subjects? If they are mainly of the combined subtype, it might be worth (see (6)) to repeat (all) classification analyses just with the combined-type subjects.

See response 2.5 for more information on the independent group. See response 1.4 for the subtype bias.
As for the gender bias we found very little difference between ADHD girls and boys or between Control girls and boys. Therefore gender does not affect the classifiers in a significant manner. The ROC curve results have been placed in a new table in the paper. The corresponding ROC curves are shown here in figure 3.

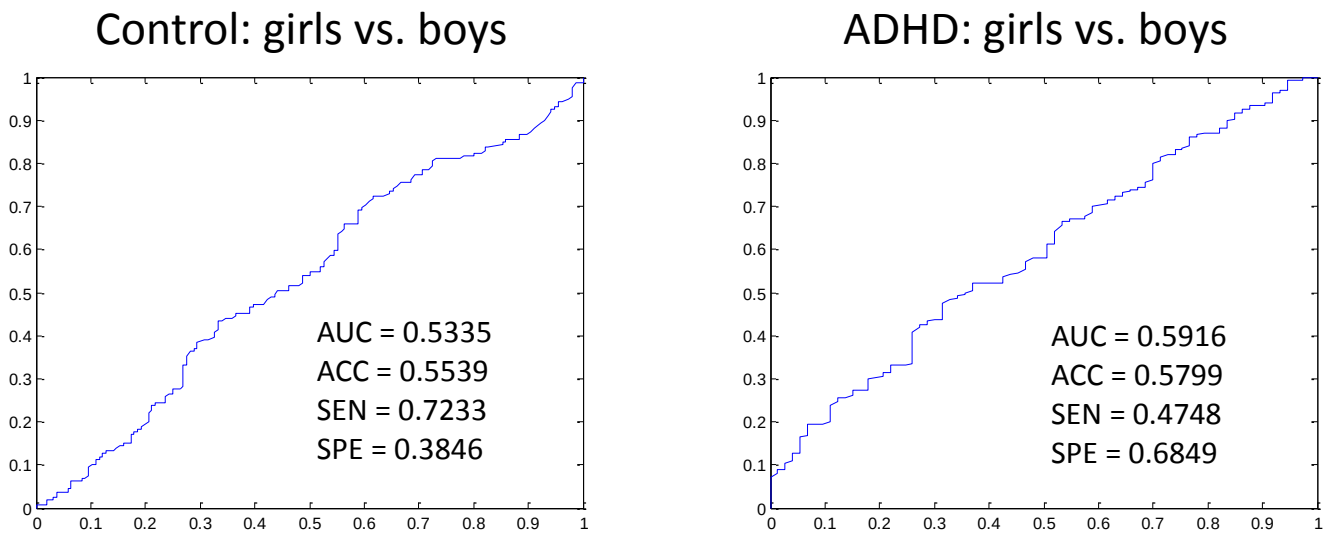


Figure 3. The ROC curve for the separation of Control girls vs. Control boys (left) and ADHD girls vs. ADHD boys (right) based on the ADHD vs. control index. The AUC value is below 0.6 and thus there is very little separation.

1.6) I found it surprising that 22% of the subjects had two or more comorbid disorders. What type? Which group was dominantly affected, e.g. the combined subtype? Due to the large number of subjects, the authors could repeat the classification analysis removing this 22% of data.

Response: See general response, we want both to be included. No difference in ADHD clean vs. ADHD with comorbidity was found using the ADHD vs control index (see figure 4). Therefore ADHD subtype does not affect the classifiers. The ROC curve results have been placed in a new table in the paper. The corresponding ROC curve is shown here in figure 4.

According to Patel et al. (2012) 67% of children with ADHD have at least one mental health disorder or neurodevelopmental disorder and 34% have two or more. Our dataset is not far from this statistic. The cohort should reflect the typical cross section of the individuals who visit the clinic for ADHD diagnosis in order to develop a clinical diagnostic method for the general clinical population. Information on how each subgroup of ADHD is affected is in table 1. The types of comorbidities and their frequency are shown in table 3.

Patel N, Patel M, Patel H. ADHD and Comorbid Conditions. In: Norvilitis JM, editor. Current Directions in ADHD and Its Treatment [Internet]. InTech; 2012 [cited 2014 Oct 14]. Available from: http://www.intechopen.com/books/current-directions-in-adhd-and-its-treatment/adhd-and-comorbidity

Table 3. Comorbid disorders and their frequency

ODD	92
Autism	41
Anxiety	28
Tics	25
Depression	11
OCD	4
Other1	75
Other2	16

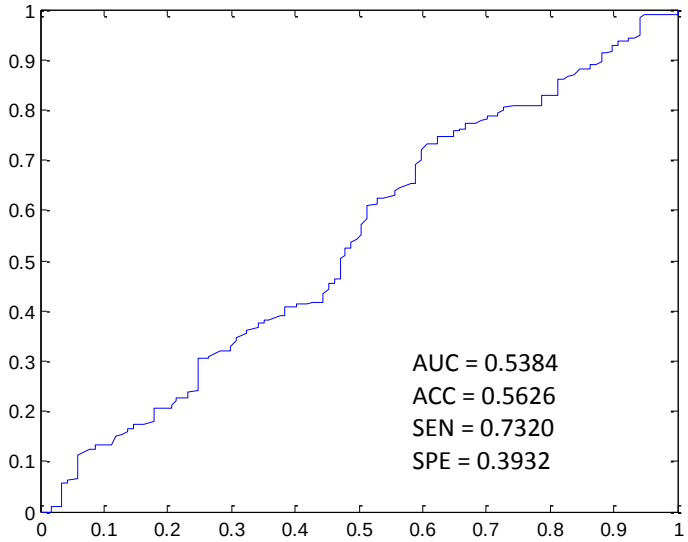


Figure 4. The ROC curve for the separation of ADHD comorbidity vs. no comorbidity based on the ADHD vs. control index. The AUC value is close to 0.5 and thus there is no separation.

1.7) Introduction: There are a bunch of classification studies using power values apart from Magee et al (2005). In this paper, specificity was only 40% but other papers (e.g. Poil et al., 2014) found much better classification rates (> 70%). Authors could cite other studies, using classification analyses, to balance the introduction.

Response: We have improved the references in the submission:
"Magee et al (20) used a combination of factor, cluster and regression analysis to develop a diagnostic classifier for a homogeneous group of ADHD patients based on EEG power measures across all frequency bands and channels, obtaining an overall classification sensitivity of 89.0% and a specificity of 79.6%. Poil et al. (21) classified ADHD adults vs. controls with 67% sensitivity and 83% accuracy with support vector machine classification. Such a multivariate statistical approach can be further extended to include multiple EEG features, including coherence measures (22,23)."

1.8) MAJOR: So far "coherence" or "coherence features" (page 16) remain loose definitions. The authors need to review some of the existing (ADHD) coherence studies in more detail (e.g. papers by Dupuy). What kind of coherence measures were used in those studies (references 21 and 22 but only in other studies)? Which (scalp/brain) regions demonstrate abnormal connectivity in ADHD? Critically, what kind of coherence (for which frequency band) was used

in the present study (lagged coherence, zero-lag coherence, global synchronization)??? The information provided on page 9 is not sufficient.

Response:
It is beyond the scope of this work to analyse specifically the properties of the qEEG features and the coherences. The spirit of the feature extraction is to capture the relevant degrees of freedom in the multivariate signal.

The above text was added to the paper.

In order to capture the connectivity degrees of freedom we chose to consider the autocorrelation function between electrodes in the average montage. The spectral features related to connectivity were then estimated from that. In practice this is done by considering the autocorrelation function for 2sec segments and evaluate the spectrum for each segment. The segments considered are all segments within the selected recording with 1sec overlap. A Bartlet window is applied to each segment. The analysis results in a spectrum for each segment. This ensemble is then used in order to estimate a representative spectrum by applying robust fitting over all the spectrums. In that way incidental artefacts are avoided. The classical qEEG spectral features for each channel are obtained in a similar manner, again in the average montage.

1.9) Although the ADHD rating scale IV is a validated scale, the rationale of using a "score of 1.5 standard deviations above the age-appropriate norm" needs to be justified.

Response: The 1.5 SD cut off is widely used in screening in clinical practice to determine which children need to undergo a more thorough clinical evaluation for ADHD. This cutoff may seem rather high for the normal comparison group but the rationale was that we did not want a "super-healthy" group but rather a mixed group such as might be seen in a referred group of children in normal clinical work.

This response also applies to 2.3

1.10) Methods: How much clean (after artefact correction) EEG data was finally obtained in the examined groups? This is important to know since the overall recording duration was only three minutes.

Response:
See response 1.8. An automatic artefact removal scheme is applied by a robust fit of each feature.

1.11) Major 1: I am missing pure coherence results. Do the groups (especially in the four age sub-groups) differ in coherence in specific frequency-bands? So far, it is unclear whether the classification results really support the findings by Magee et al (2005) and Barry et al. (2002). See also comment (15).

Response: The strongest features that appear in the classifiers represent inter-hemispheric coherences in the central regions (C3/C4 and T3/T4). This is in accordance with the results of Barry et al. (2002). This has been accounted for in the paper. The details can be found in the paper.

1.12) Major 2: The authors have data on power values as well (Table 3). What was the reason to not include them into the construction of the classifiers? Would this not lead to better classification results? In general it would be interesting to see if power-related classification results differ to coherence-related classification results.

Response: See response 1.8.

1.13) Why the authors did use nine (age-independent) classifiers for the construction (page 10, bottom)?

Response: The methodology is based on our experience from our work with the elderly and diagnosis of dementia. The point is to avoid imbalances in group sizes and increase robustness. Also, the time it takes to calculate the AUC for thousands of classifiers is relatively short (days) for group sizes of 100.

1.14) Can the authors discuss the results of Table 5 (and Table 7) in more detail, as the highest accuracy and AUC values was achieved for age group 4? Could this be the result of a developmental lag, which is present in ADHD population (whereas the non-ADHD subjects already moved to puberty)?

Response:
According to Shaw (2007) the cortical development of children with ADHD is lagged behind those not diagnosed with ADHD. As the oldest age group is close to entering puberty, it is possible that a bigger part of the subjects in the control group has already moved to puberty than in the ADHD group. As children reach puberty they experience hormonal and physical changes and this might be one of the factors that explains the high accuracy in the oldest age group.

Shaw P, Eckstrand K, Sharp W, Blumenthal J, Lerch JP, Greenstein D, et al. Attention-deficit/hyperactivity disorder is characterized by a delay in cortical maturation. Proc Natl Acad Sci U S A. 2007;104(49):19649–54.

1.15) Also the results are not discussed in the context of other EEG coherence studies (to ADHD).

Response: See responce 1.11.

1.16) Finally, a clinician will not necessarily know what to do with these results. First, the classification results are moderate (< 80%). Second, the results are averaged across different sub-types of ADHD, and third, medication was not talking into account so far. As already written, authors can address issues (2) and (3) by more detailed analysis.

Response: Clinically the diagnosis of ADHD is based on several sources of information, amongst them standardized interviews, rating scales, developmental and medical history. No single source of information can be expected to have 100% accuracy so a clinical decision must be based on several sources, sometimes providing conflicting information. The EEG classification comes into this as an additional source of information and for those of the group of authors who have already used this in their clinical work it has proved to be a helpful addition.

Also, see the general response.

1.17) Although a nice try, authors do not have any support for the assumption that their connectivity results are driven/linked by alterations in the neurotransmitter (e.g. dopaminergic) system. In fact, results could be partially driven medication, which affects neurotransmission, and ultimately coherence.

Response:

See general response and response 1.3, there is no difference in ADHD medicated vs. ADHD not medicated based on the ADHD vs. Control index (see figure 1).

The following change was made to the paper.

"In addition, our findings are consistent with the increasingly compelling results linking ADHD to deficits in brain connectivity, ~~driven by the neurotransmitter systems, in this case dopaminergic and noradrenergic systems~~ (7,37,38)."

1.18) Only because all recordings were made in Iceland, does not allow saying: "genetics…were quite tightly controlled". Please rephrase (e.g. page 16).

Response:
The following change was made to the paper.

Hence both the underlying population ~~(in terms of genetics and environment)~~ and EEG equipment were quite tightly controlled.

Reviewer: 2
Reviewer Name Sandra Loo and Iman M. Rezazadeh
Institution and Country, University of California, Los Angeles, USA
Please state any competing interests or state 'None declared': None declared

I'm not sure about this. Statistical pattern recognition is used to determine the classifier. I am not an expert in the use of this algorithm.

The goal of this study is to develop a multivariate EEG biomarker for the diagnosis of ADHD. The authors report 72-79% diagnostic accuracy for ADHD and conclude this may be useful for improving accuracy of initial diagnosis and ongoing treatment monitoring of children with ADHD.

Overall, there are several strengths of this manuscript. The paper is clearly written and includes a very large sample. The use of EEG coherence is novel and sophisticated statistical techniques (statistical pattern recognition; SPR) are used in determining the classifier. Finally, an independent sample of EEGs was used to test the classifier.

There are several weaknesses, however, particularly with regard to study methodology, data analysis, and scientific significance that diminish the impact of this manuscript. The lack of detail makes it difficult to review the study's methodological rigor and the omission of the measures that are retained for the final classifier makes evaluation of the scientific contribution and impact impossible. Specific details that would make the report of this study more complete and comprehensive follow.

2.1) With regard to study methodology, more detail and better characterization of the ADHD sample is needed. The only exclusion criterion for the ADHD sample was 'moderate to severe intellectual disability', what IQ range does this include? Was the intelligence measured in the control sample and the same exclusion criterion applied?

Response:
IQ tests were not conducted for the normal comparison group but moderate or severe intellectual disability was excluded by parent information about mental or developmental problems. Tests of intellectual level were available in the two clinics from which the ADHD participants were recruited, making it possible to exclude children with moderate and severe intellectual disability. IQ scores were not recorded in the dataset of the study.

For more details on the ADHD sample see response 1.1, 1.3, 1.4, 1.5 and 1.6

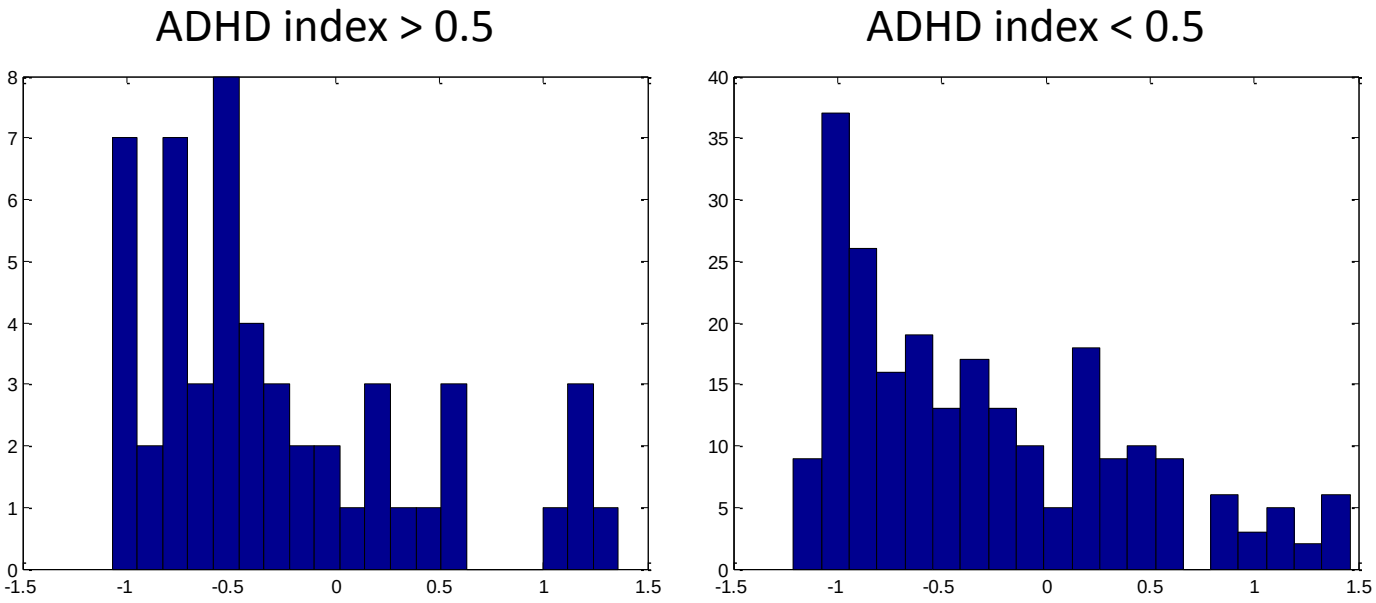
2.2) Further, it is noted that patients with co-morbid autism spectrum disorder (ASD) were included and reported to be a frequent co-morbidity. All co-morbidities and their frequencies should be reported in the paper. Because ASD is often so much more severe than ADHD, it is questionable whether subjects with ASD should have been included in the sample. The authors should report whether ASD or any other co-morbidity affected the diagnostic accuracy of the classifier.

Response: See general response and response 1.6.

2.3) The control sample seems to have undergone minimal evaluation. The threshold of <1.5 SD below the age norm seems too liberal for a control sample and may have included controls who have sub-threshold ADHD. Authors should report behaviour rating scores for control subjects who were misclassified in the analysis.

Response: See general response and response 1.9.

Figure 5. A histogram of the ADHD rating score in numbers of standard deviation from norm. The misclassified part of the Control group is on the left and those correctly classified on the



right. The misclassified individuals do not have rating scores that are obviously different from those who are correctly classified.

There is no difference in the number of standard deviations from normal ADHD rating scores between the individuals in the Control group who are misclassified compared to those who are correctly classified.

2.4) Overall, more demographic (SES, race, gender) and clinical (comorbidities, ADHD severity and symptoms, IQ, behavioural functioning) data should be reported for both samples, ADHD and control, so that generalizability of the data can be assessed.

See response 1.1. We did not collect data on SES of the subjects. The National Bioethics Committee did not give permission to collect data on race. In Iceland a person can be identified by the month of birth and race, if the race is not Caucasian.

Data processing and analysis:

2.5) Because there were 295 ADHD patients and 350 recordings, there are some recordings that were made on the same person during on and off medication conditions, as explained in the manuscript. Presumably, these are considered 'independent' even though recordings from the same person were likely included in the training and testing cohorts. Since the

testing cohort only includes 35 EEGs, it could very well be that all of the EEGs were from people also used in the training phase. This calls into question just how 'independent' the testing cohort really was. Although there may have been some EEG changes while on and off medications, the EEGs are likely highly correlated and unlikely to be completely independent in the training and testing cohorts. This contradicts the authors statements that the testing of the classifier was done in an independent sample.

Response: We have replaced the individuals in the independent group who were also in the training cohort (in a different medication state). The replacements are taken from a group of ADHD individuals who have been recruited since the article was first submitted.

Table 4a. New independent group

		Total	male	female	age	Comorbit, one	Comorbid, 2 or more	ADHD med	ADHD rating SD from M
ADHD group		**36**	**26**	**10**	**10,53**	**20**	**5**	**10**	**2,72**
	ia	13	9	4	11,47	9	3	4	2,73
	com	22	16	6	10,09	11	2	6	2,99
	adhd hyp	1	1	0	13,95	0	0	0	2,13
controls		**36**	**19**	**17**	9,65	**0**	**0**	**0**	**-0,46**

Table 4b. Old independent group

		Total	male	female	age	Comorbit, one	Comorbid, 2 or more	ADHD med	ADHD rating SD from M
ADHD group		**35**	**24**	**11**	**9,87**	**20**	**4**	**14**	**3,06**
	ia	14	11	3	10,49	9	3	8	2,34
	com	21	13	8	9,47	11	1	6	3,53
controls		**36**	**19**	**17**	**9,65**	**0**	**0**	**0**	**-0,46**

2.6) It is not clear how authors dealt with the variation among the recording sites, even though similar systems and settings were used, before combining the EEG data and training the classifier. More details should be given regarding how the EEG data were recorded and processed. Please provide information on how the data were cleaned of artifacts and noise. What was the minimum number of epochs needed to be included in the analysis?

Response: Variation among recording sites/clinics is the fact of life for clinicians. In the spirit of the cross sectional approach we dealt with the variation by not doing anything about it. See response 1.8 for the other details.

2.7) Were 100% of the EEGs obtained on the ADHD sample usable and included in the analysis?

Response: According to the protocol the recording was to be repeated if a technical problem occurred. As a result, no subjects were excluded in terms of bad quality of the EEG.

2.8) How were the power spectrum, relative power and coherence calculated?

Response: See above. See response 1.8.

2.9) How was volume conductance accounted for when calculating coherence?

Response: Not relevant when following the above procedure. See response 1.8.

2.10) For the statistical analysis, a more complete description of the rationale of the SPR classification method/classifier should be given; Why were other widely used classifiers like Support Vector Machine, ANN, and ANFIS not used?

Response: The Statistical Pattern Recognition is based on SVM. We have clarified this in the paper.

2.11) The 'training' set has near 90% of the data, which is more than what is normally suggested for training a classifier. Please give the rationale for this rather than using more standard percentages, e.g., splitting the dataset around 70% 'training set', 20% ' PubMed test set' and 10% 'validation set'.

Response: The statistical properties of each classifier are estimated using 10-fold cross validation. 10-fold cross validation is common, and very likely the most common of all by far. It has been labelled "the industrial standard". We also have an independent validation set. This means that the split is roughly: training set 80%, cross-validation set 10% and independent validation set 10%, which is close to the split mentioned by the reviewer. We list two references which support our choice of the value k=10 in k-fold cross validation (note the above mentioned split is close to 3-fold cross validation).

Knafl GJ, Grey M. Factor analysis model evaluation through likelihood cross-validation. Stat Methods Med Res 2007;16:77-102.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Mellish CS ed. Proceedings of the 14th international joint conference on artificial intelligence. Morgan Kaufman; 1995: 1137–43.

2.12) What is the main reason/s that accuracy of age-dependent /independent classifiers differs? Is this related to sample size or other factors?

Response: The difference is probably due to the large age range in case of the age-independent classifiers. Puberty could also be an important factor, see response 1.14.

2.13) What is the mutual information/cross-correlation of the coherence features that are used in the analysis?

Response: See response 1.8.

2.14) It is not clear in the paper why subsets of 20-features have been chosen for the evolutionary classification method? Why 20? How was the redundancy among the 20-set of features reduced?

Response: 20 features were chosen because we have experience using that number of features from our work with the elderly and diagnosis of dementia where the size of groups is the same (around 100). Therefore, even though some features are redundant, we have obtained a usable set of features. As we have explained in response 1.11, we do not consider the 20 features of the optimal set of features as the 20 best features that can separate the ADHD and Control group. We explain in the paper how we can estimate the most relevant features.

2.15) What is the final subset of features used in the classifiers? Please list the features.

Response: See response 1.11.

2.16) Did the authors test the association between the final features and clinical features of the sample? This is an important step to determine the functional significance of the features used in the classifier.

Response: The ADHD vs control index does not correlate with the ADHD RS for the ADHD group in the training set (the correlation coefficient is 0.033).

2.17) The omission of the results also makes it impossible to compare to previous studies that have examined coherence measures in ADHD as well as previous findings on the neurobiology of ADHD. Diagnostic accuracy of ~80% leaves approximately 20% of subjects incorrectly diagnosed. Authors should outline the appropriate usage of an algorithm with only moderate diagnostic accuracy.

Response: See response 1.16.

Reviewer: 3

Reviewer Name Victoria Harris
King's College London, United Kingdom
Please state any competing interests or state 'None declared': None

Comments:
This is primarily a statistical review. The paper explores an interesting application of classification tools to EEG data in order to identify individuals with ADHD. The authors use ROC analysis, comparing sensitivity, specificity and area under the curve, to assess the performance of age specific and non-age specific classifiers. The research is novel in that although similar methods have been applied successfully to other disorders, this is the first study for which these methods have been applied to a group with ADHD. This suggests that this technique may have application as a diagnostic tool for clinicians. Overall the analysis appears sound but it would be useful to the reader for the authors to supply additional details on the classifiers and how they were constructed.

3.1) The authors mention that they divide their data in to a training cohort of 315 cases and the same number of controls and a test cohort of 35 cases and 36 controls. They mention that 350 of the case observations were obtained from 295 individuals. Were any of the individuals that were measured more than once included in both the test and the training set and if so did you account for this?

Response : See response 2.5

On page 10, lines 4-11 the authors describe how they selected their classifiers using statistical pattern recognition and an evolutionary algorithm based on the target of the AUC. Statistical pattern recognition refers to a set of statistical learning techniques, in this case for the purpose of classification.

3.2 It would be helpful to have further detail here of the SPR technique used to allow the reader to better understand how the candidate classifiers have been constructed.

Response: We have added to the methods section to account for this.

3.3 An evolutionary algorithm is an optimisation technique that models candidate solutions as individuals in a population, modelling recombination of features within each iteration, and here is used to select a set of twenty features associated with each classifier. It would be beneficial to add a brief description of the method and some additional references.

Response: We have added to the methods section to account for this.

We feel that a detailed description of the evolutionary algorithm in the paper is not important or of interest to the general reader. Therefore we include the details here instead.

In the evolutionary scheme an individual is a classifier constructed from 20 features and each feature is a single gene in the gene pool. The evolutionary algorithm selects the 20 best individuals, based on the classifiers' AUC value, as parents for the following generation. The first generation of parents is selected at random from the gene pool. Around 200 offspring are created during each evolutionary cycle by predefined mating and mutation schemes. In the mating scheme two parents are selected at random, and then a random number of genes are interchanged to create two individuals. This is repeated 100 times and redundant individuals removed. In the mutation scheme an individual is chosen at random from the parents or new-born children and 5, 10, or 15 (at random) genes replaced by randomly chosen genes. Up to 120 individuals are created in this manner, depending on the number of unique children born in the cycle. This is repeated until no further improvement in AUC value is evident. Normally it

is enough to run 100 cycles and the number of classifiers constructed ranges from 5000 – 7000 during those cycles.

3.4) In the discussion the authors mention the possible issue of confounding due to gender. They also mention that a large proportion (66%) had comorbid disorders. The authors also mention that SPR of EEG data has previously been successfully used for data on autism. How do the classifiers for autism found in previous studies compare with those for ADHD for in this study?

Response:
Duffy and Als (2012) present a phenotype of ASD and the result is a complex pattern of EEG coherences. The main features in the ASD results, reduced coherence in the left temporal-frontal regions, are different from our main findings, elevated interhemispheric coherences in total power and all frequency bands but α2 band in the central region of the brain.

Duffy FH, Als H. A stable pattern of EEG spectral coherence distinguishes children with autism from neuro-typical controls - a large case control study. BMC Med. 2012 Jun 26;10(1):64.

3.5) Similarly has any previous work explored gender differences in EEG readings?

In this study the ADHD vs Control index is independent on gender, see response 1.5 and the differences in ratio are not likely to have an effect on our results. Therefore we took out the comment on the effect of gender in the discussion.

~~A possible confound of the study is the significant difference in gender balance between the clinical group and the control group. The male:female ratio is even in the control group, but in the clinical group it is 3:1. The difference is not unexpected, as ADHD is known to be more commonly diagnosed in boys (41).~~

VERSION 2 – REVIEW

REVIEWER	Lars Michels Institute of Neuroradiology University Hospital of Zurich
REVIEW RETURNED	21-Nov-2014

GENERAL COMMENTS	I think the authors have carefully most of the major and minor concerns. Few, final adaptations might further improve the paper. 1) Maybe it would be good to briefly describe the results of the four types of mixed types (see general response on page 25). "We calculated how well the four types of mixed groups separate based on the indices obtained from the age independent ADHD vs control classifiers. In short, there was no separation in any one of them and the index may be considered independent of those factors." 2) It is good that the authors also edit information on medication to Table 2. 3) The authors could mention that type and frequency of comorbidities is not far away from the Patel (2012) study.

	4) Please mention a reference (if possible) for the 1.5 SD cut off (issue 1.9). 5) Issue 1.8: Move the response to the text and mention the total nr of included segments (see issue 1.10). 6) Issue 1.13: Please mention this in the text.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name Lars Michels
Institution and Country Institute of Neuroradiology
University Hospital of Zurich

The authors have carefully most of the major and minor concerns. Few, final adaptations might further improve the paper, which I have attached in a separate file.

Congratulation on this nice work.

I think the authors have carefully most of the major and minor concerns. Few, final adaptations might further improve the paper.

1) Maybe it would be good to briefly describe the results of the four types of mixed types (see general response on page 25).

Response: This has already been added to the manuscript, see *Table 6*.

"We calculated how well the four types of mixed groups separate based on the indices obtained from the age independent ADHD vs. control classifiers. In short, there was no separation in any one of them and the index may be considered independent of those factors."

2) It is good that the authors also edit information on medication to Table 2.

Response: In *table 1* we describe the ADHD subjects in terms on if they were medicated or not at the time of the recording. In the *Method* section, *EEG Data Acquisition* we list the type of medication.

3) The authors could mention that type and frequency of comorbidities is not far away from the Patel (2012) study.

Response: We have included following in the section *Method, Participants:*

These types and frequencies are comparable to the comorbidities presented in the study of Patel (29).

We also added a new **Reference nr. 29 (Patel et. al)**

4) Please mention a reference (if possible) for the 1.5 SD cut off (issue 1.9).

Response:

The cut off is chosen according to clinical experience. For a discussion of different cut offs see the manual of the ADHD Rating Scale (30). For further information we have included a local study on the use of ADHD rating scale in Icelandic school children to the reference list. (32).

30. DuPaul GJ, Power TJ, Anastopoulos AD, Barry RJ. ADHD Rating Scale--IV (for Children and Adolescents): Checklists, Norms, and Clinical Interpretation. New York: The Guilford Press; 1998.

31. Zhang S, Faries DE, Vowles M, Michelson D. ADHD Rating Scale IV: psychometric properties from a multinational study as a clinician-administered instrument. Int J Methods Psychiatr Res. 2005;14(4):186–201.

32. Magnússon P, Smári J, Grétarsdóttir H, Prándardóttir H. Attention-Deficit/Hyperactivity symptoms in Icelandic schoolchildren: assessment with the Attention Deficit/Hyperactivity Rating Scale-IV. Scand J Psychol. 1999 Dec;40(4):301–6.

5) Issue 1.8: Move the response to the text and mention the total nr of included segments (see issue 1.10).

Response: We have included following in the section *Method, Signal Processing and Statistical Analysis:*

It is beyond the scope of this work to analyse specifically the properties of the qEEG features and the coherences. The spirit of the feature extraction is to capture the relevant degrees of freedom in the multivariate signal. **In order to capture the connectivity degrees of freedom the choice was to consider the autocorrelation function between electrodes in the average montage. The spectral features related to connectivity were then estimated from that. In practice this is done by considering the autocorrelation function for 2sec segments and evaluate the spectrum for each segment. The segments considered are all segments within the selected recording with 1 sec overlap. A Bartlet window is applied to each segment. The analysis results in a spectrum for each segment. This ensemble is then used in order to estimate a representative spectrum by applying robust fitting over all spectra. In that way incidental artefacts are avoided. This evaluation was repeated for five consecutive 36 sec intervals of the recording. The outcome for the intervals were then averaged and applied. The classical qEEG spectral features for each channel are obtained in a similar manner, again in the average montage.**

6) Issue 1.13: Please mention this in the text.

Response: We have included following in the section *Method, Signal Processing and Statistical Analysis:*

This methodology is based on our experience from our previous work (23). The point is to avoid imbalances in group sizes and increase robustness. Also, the time it takes to calculate the AUC for thousands of classifiers is relatively short (days) for group sizes of around 100.